MONOCULAR SLAM SUPPORTED OBJECT RECOGNITION

Sudeep Pillai & John J. Leonard
Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Robots equipped with a single RGB camera need to continuously recognize and localize all potential objects in its immediate environment.

**Single RGB Camera**
- Versatile

**Multi-view Object Detection**
- Camera & Object localization by leveraging SLAM

**Robust**
- Avoid spurious detection/mis-classification

**Scalable**
- Runtime sub-linear in categories identifiable
PRIOR WORK

• Shift in Object Recognition
  
  PAST
  • DPM and Sliding Window Detection
  • Category-specific templates
  
  TODAY
  • Most state-of-the-art object recognition methods today utilize object proposals

R-CNN: Regions with CNN features
Girshick et al. (CVPR 2014)
PRIOR WORK

• Advances in Feature Encoding
  • Dense SIFT + BoVW / VLAD / FV
  • Pros: Scalable, category-agnostic
  • Cons: Lack object localization

• BoVW methods + object proposals with box encoding methods

PAST

TODAY

Fisher and VLAD with FLAIR
van de Sande et al. (CVPR 2014)
**PRIOR WORK**

- New map representations
  - Leverage semi-dense maps
  - Edges that are consistent across views could potentially propose objects
KEY CONCEPT

SLAM-capable robots equipped with a single RGB camera need to continuously recognize and localize all potential objects in its immediate environment.

**Single RGB Camera**
- Monocular SLAM supports improved recognition
- (Semi-Dense Mapping Backend)

**Multi-view Object Detection**
- Objects easily tease apart to enable better proposals
- (Proposals from Semi-Dense Maps)

**Robust**
- Reduced false positives via view correspondence from SLAM
- (Multi-view prediction)

**Scalable**
- ~1.6-1.7 s for 5-50 object categories
- (FLAIR encoding)

Semi-Dense Monocular Reconstruction with Object Labels
MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

SEMI-DENSE MAPPING

As a camera-equipped robot moves around in its immediate environment, the camera is localized and simultaneously a semi-dense map is constructed.
MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

MULTI-VIEW OBJECT PROPOSALS

Scale-ambiguous Reconstruction

Filtered Reconstruction

Semi-Dense Depth Filtering
SVO: Forster et al. (ICRA 2014)

ORB-SLAM
Mur-Artal et al.

Low-density region pruning

The semi-dense reconstruction provide spatio-temporally consistent edges, a strong indication of object presence.
MONOCULAR SLAM-SUPPORTED OBJECT RECOGNITION

MULTI-VIEW OBJECT PROPOSALS

Scale-ambiguous Reconstruction

Multi-scale Over-segmentation

Semi-Dense Depth Filtering
SVO: Forster et al. (ICRA 2014)

ORB-SLAM
Mur-Artal et al.

Density-based segmentation over 4 spatial scales

Object proposals are extracted from the reconstructed scene via a multi-scale density segmentation step, and are further refined for occlusions.
MONOCULAR SLAM-SUPPORTED
OBJECT RECOGNITION

MULTI-VIEW OBJECT PROPOSALS

Multi-scale Over-segmentation

View-Consistent Proposals

Density-based segmentation over
4 spatial scales

Candidate proposals projection

The segmented regions are projected onto each keyframe enabling view-consistent
object proposals that perform better than frame-based (BING) proposal methods
The object proposals in each frame are encoded (via FLAIR) and subsequently classified, before their evidence is probabilistically fused across all keyframes.
SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION

SLAM-Supported Recognition (Ours)

Frame-based Recognition (Classical approach)
Object hypotheses are aggregated across views resulting in correct classification.

SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION

SLAM-Supported Recognition (Ours)  Frame-based Recognition (Classical approach)
SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION

SLAM-Supported Recognition  
(Ours)

Frame-based Recognition  
(Classical approach)
SLAM-SUPPORTED vs. FRAME-BASED OBJECT RECOGNITION PERFORMANCE

- **Recognition performance**
  - RGB: Superior to frame-based methods
  - RGB-D: Comparable to state-of-the-art SLAM-based recognition methods
  - Improved and robust performance compared to frame-based methods via SLAM-based view aggregation

Detection-based object labeling in 3D scenes
DetOnly - Lai et al. (ICRA 2012)

Unsupervised feature learning for 3D scene labeling
HMP2D+3D, Det3DMRF - Lai et al. (ICRA 2014)

**SINGLE-VIEW RECOGNITION**

- Ours (RGB): 81.5
- DetOnly (RGB): 61.7

**MULTI-VIEW RECOGNITION**

- Ours (RGB): 89.8
- Det3DMRF (RGB-D): 91.0
- HMP2D+3D (RGB-D): 90.9

mAP - (mean Average Precision)
**SLAM-SUPPORTED OBJECT PROPOSALS & SCALABILITY**

- **Multi-view Object Proposals**
  - Consistent object hypothesis supported by newer semi-dense map representations
  - Achieve similar Recall-vs-IoU with fewer object proposals

- **Scalable Runtime Performance**
  - Category-agnostic BoVW + FLAIR encoding
  - Run-time nearly independent of object categories identifiable

---

### MULTI-VIEW OBJECT PROPOSAL PERFORMANCE

<table>
<thead>
<tr>
<th>IoU Threshold</th>
<th>Multi-view (12.9)</th>
<th>BING (28.0)</th>
<th>BING (81.7)</th>
<th>BING (157.8)</th>
<th>BING (254.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>1.0</td>
<td>0.85</td>
<td>0.75</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>0.4</td>
<td>0.95</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9</td>
<td>0.85</td>
<td>0.75</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>0.6</td>
<td>0.85</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>0.7</td>
<td>0.8</td>
<td>0.75</td>
<td>0.65</td>
<td>0.55</td>
<td>0.45</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7</td>
<td>0.65</td>
<td>0.55</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.55</td>
<td>0.45</td>
<td>0.35</td>
<td>0.25</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>0.45</td>
<td>0.35</td>
<td>0.25</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**IoU** - Intersection Over Union  
(#{}) - Number of object proposals

### RUNTIME PERFORMANCE

<table>
<thead>
<tr>
<th>Runtime (s)</th>
<th>Ours</th>
<th>DetOnly</th>
<th>Ours</th>
<th>DetOnly</th>
<th>Ours HMP2D+3D (#{})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6 s</td>
<td>Ours</td>
<td>DetOnly</td>
<td>Ours</td>
<td>DetOnly</td>
<td>(5) 1.6 s</td>
</tr>
<tr>
<td>1.6 s</td>
<td>Ours</td>
<td>DetOnly</td>
<td>Ours</td>
<td>DetOnly</td>
<td>(10) 1.7 s</td>
</tr>
<tr>
<td>1.7 s</td>
<td>Ours</td>
<td>DetOnly</td>
<td>Ours</td>
<td>DetOnly</td>
<td>(51) 1.8 s</td>
</tr>
<tr>
<td>1.8 s</td>
<td>Ours</td>
<td>DetOnly</td>
<td>Ours</td>
<td>DetOnly</td>
<td>(5) 1.6 s</td>
</tr>
</tbody>
</table>

Lower is better

(#{}) - Number of categories identifiable
• Developed a **SLAM-supported object recognition system**, that is able to provide accurate, robust and scalable recognition performance using a single RGB camera

• Leveraged recent advancements in mapping and feature encoding techniques to enable **view-consistent object proposals**, and an improved recognition solution whose runtime is nearly-constant to the number of objects identifiable

• Demonstrated **superior object recognition performance** compared to existing frame-based methods, and **comparable performance** to existing SLAM-based recognition methods that take advantage of RGB and depth modalities